



Data Distribution Overview

Stephen Kent, CAS Review, Mar 1, 2004

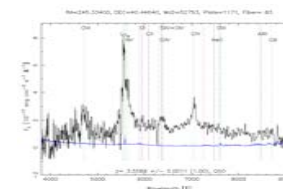
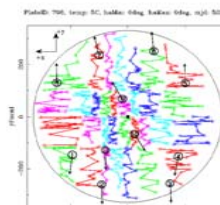
- Data processing outline
- Data distribution schedule
- Data Distribution Interfaces
 - Data Archive Server
 - Catalog Archive Server
- Data Products
 - Target vs. Best
 - Products via each interface
- Development History



Conduct of the Survey



```
run, rerun, camcol, field, id, objc type
3325, 20, 4, 108, 125, 3, 0. 0867678912151
3325, 20, 4, 108, 377, 3, 0. 0841179335687
3325, 20, 4, 108, 375, 3, 0. 0848933376756
3325, 20, 4, 108, 386, 6, 0. 0890892465128
2662, 20, 4, 283, 501, 3, 0. 0992073244309
2662, 20, 4, 283, 724, 3, 0. 0938635342479
3325, 20, 4, 108, 595, 3, 0. 1041537587318
3325, 20, 4, 108, 131, 6, 0. 1070621900713
3325, 20, 4, 109, 3, 3, 0. 115546671533759
3325, 20, 4, 108, 46, 3, 0. 11166399817049
```



**Image Sky in
5 Colors**

**Identify
10⁸ Objects**

**Select 10⁶
Galaxies and
QSOs**

**Drill 2000
Plugplates**

**Obtain 10⁶
Spectra**

**Corrected
Frames,
Atlas
Images**

**Calibrated
Object
Catalog**

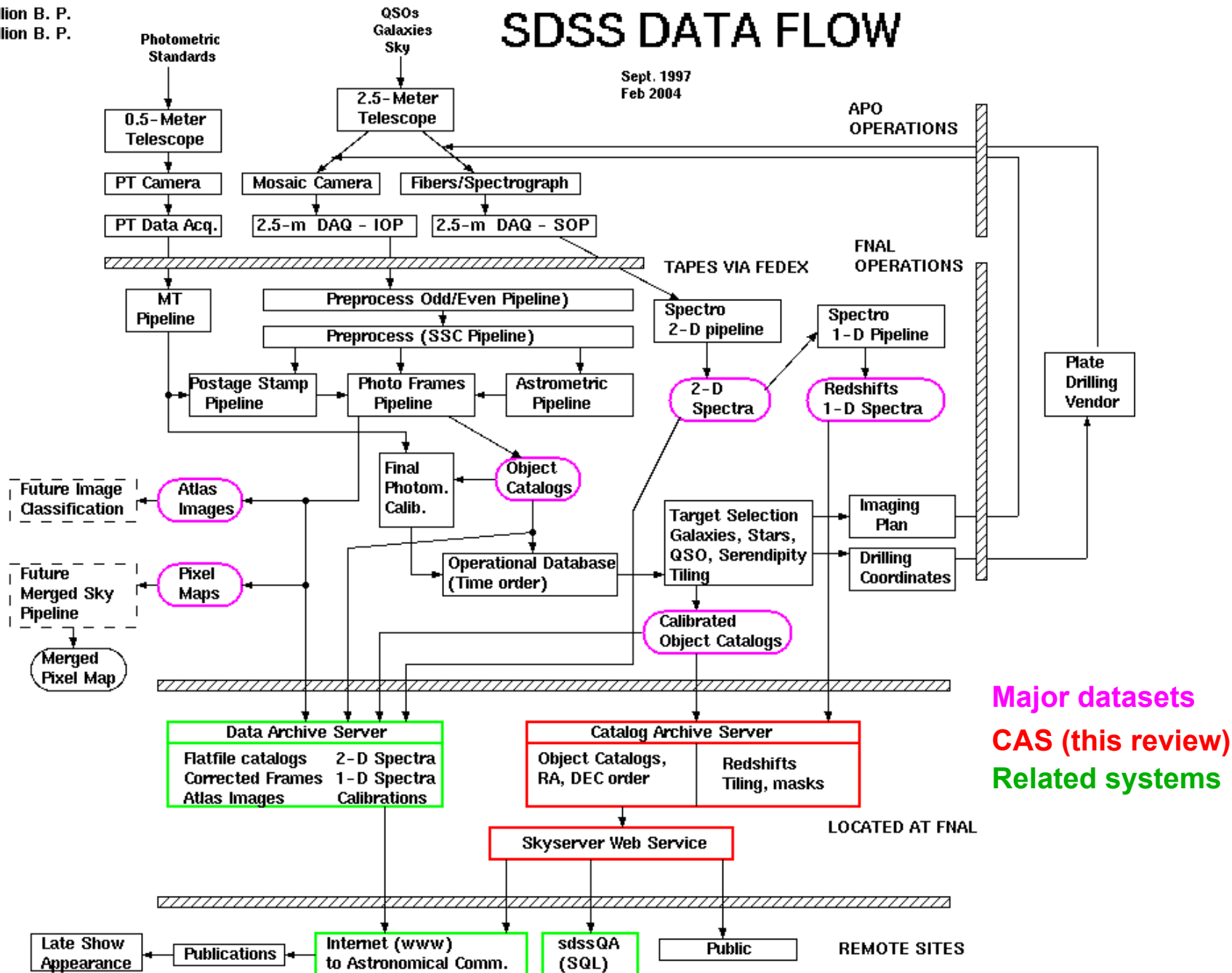
**Calibrated
Spectra,
Redshifts**



9 Billion B. P.
1 Billion B. P.

SDSS DATA FLOW

Sept. 1997
Feb 2004





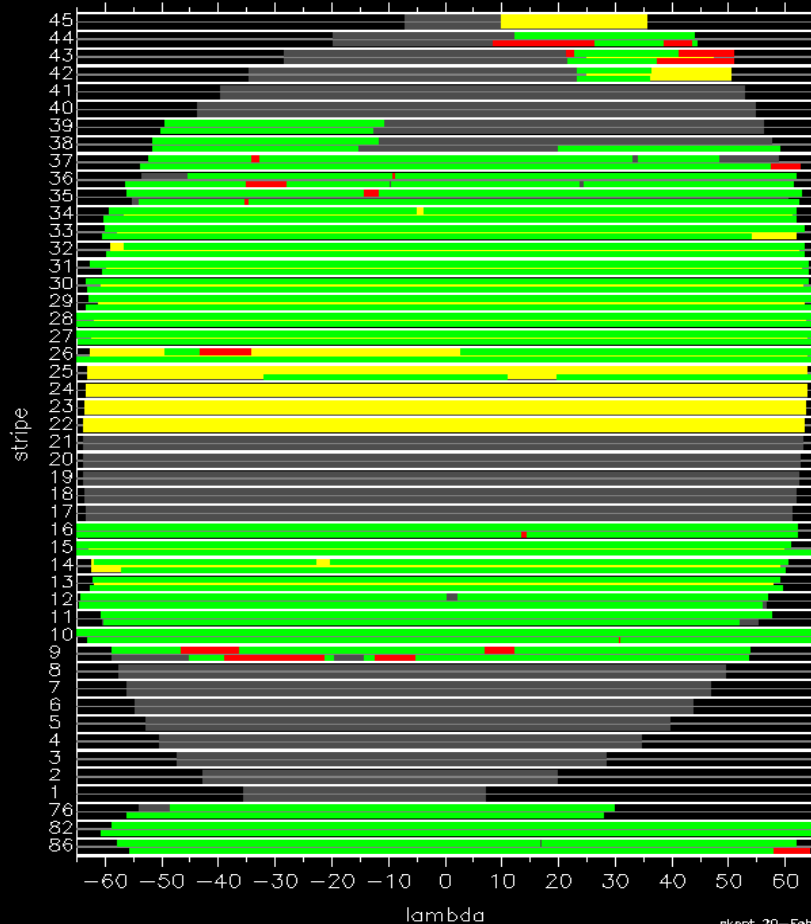
Data Collection Status

Imaging

SDSS: 2004 Feb 20

Total Area : 12898, 635mm x 1.5752
Unique Area Imaged: 6951 (82)%
(Processing Pending : 0)

Data Quality:
GOOD BASELINE
BAD



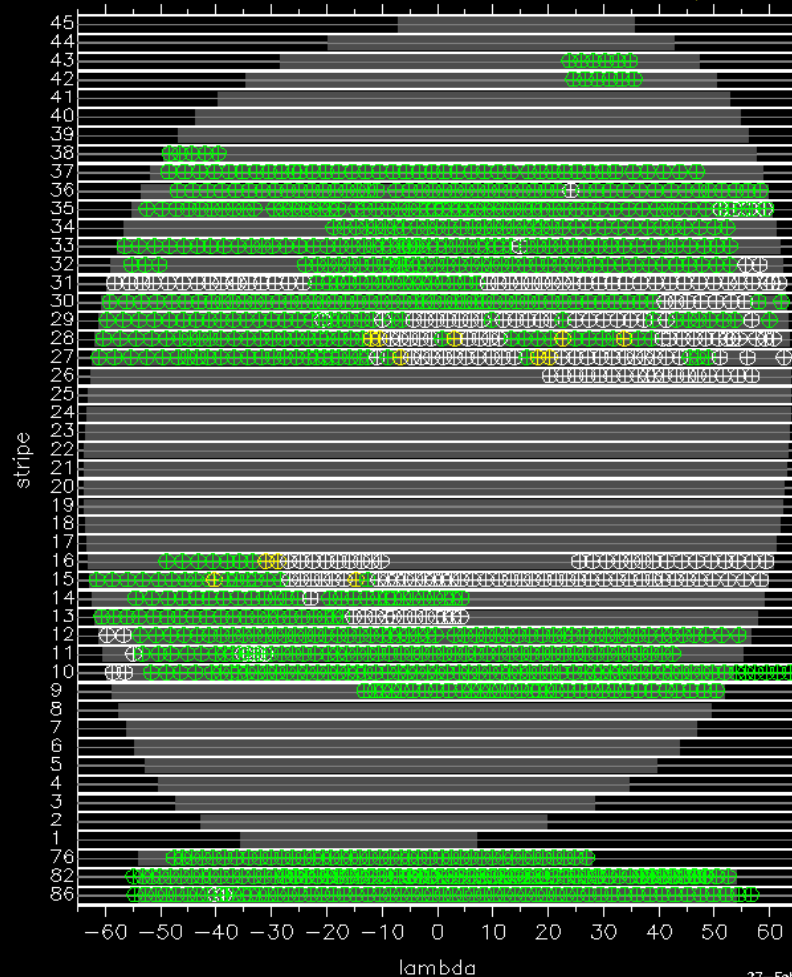
Spectroscopy

Spectroscopic Survey rerun 23

Tiles Done

Tiles defined not obs

complete





The Multiplicity Challenge

- **Data sources**
 - Imaging
 - Spectroscopy
- **Data types**
 - Pixel
 - Object catalog
- **Multiple observations**
 - Imaging overlaps
 - Spectro repeats
- **Terminology**
 - Skyserver vs. CAS
- **Access methods**
 - Flat files
 - Database
- **Imaging versions**
 - Target
 - Best (multiple)
- **Data releases**
 - DR1, DR2, etc.
- **Customers**
 - Collaboration
 - Public



Data Distribution "Customers"

- SDSS Collaboration
 - 13 institutions, including some in Germany & Japan
- Public (Astronomical community)
- *Public (general)*



Data Distribution Schedule

- Schedule for public release is defined by a plan submitted to NSF April 5, 2001

Name	Release	Photometry	Spectroscopy
Early Data Release	1-Jul-2001	5%	0%
Data Release 1	1-Jan-2003	15%	7%
Data Release 2	1-Jan-2004	47%	33%
Data Release 3	1-Oct-2004	68%	60%
Data Release 4	1-July-2005	88%	85%
Final Data Release	1-July-2006	100%	100%



Data Distribution Methods

- Data Archive Server
 - Flatfiles (pipeline outputs)
 - Fast (done in days)
 - All data accessible (all pixel data and object catalogs, all versions) (http or rsync)
 - Requires "expert" knowledge to use
- Catalog Archive Server
 - SQL Database
 - Lags data processing significantly
 - Object catalogs, spectroscopic parameters only
 - Resolve data into "seamless map of the sky"
 - Accessible via web forms or sdssQA query tool
 - Supports logical queries



What is Target vs. Best?

- **Target or Best Imaging**
 - **Specific versions of data processing, including calibrations ("rerun")**
 - **Specific "resolution" of overlapping data to form unique map of sky**
 - **Specific version of target selection code**
- **Target - used for drilling plates & obtaining spectra**
 - **Not uniform throughout the survey**
- **Best - most recent reprocessing**
 - **uniform**



Products in the CAS

- **24 Input file types**
 - **10 Imaging**
 - Lists of stars, galaxies, parameters
 - Survey geometry
 - **8 Target Selection**
 - Targeted object
 - Masks, "tiling" information
 - **6 Spectroscopy**
 - Redshifts
 - Parameters
- **63 Tables or views in database**



Products in DAS

- **18 file types**
 - **"Corrected Frames", binned frames, masks**
 - **Uncalibrated and calibrated object lists**
 - **2D, 1D spectra plus parameters**
 - **Calibrations, summary information**



Data Distribution History

- **Plan 0**

- *Push selected datasets to collaboration via network*
- *Write tapes for foreign members*
- *Archive full dataset in tape vault*
- *Abandoned due to inefficiency, excessive numbers of reprocessings, stupid tape read/write problems*
- *All collaboration members have accounts on FNAL machines; they can fetch data on their own.*



- **Plan 1 (Used for EDR)**

- **CAS implemented in Objectivity with custom SQL-like query interface (at FNAL)**
- **DAS implemented with simple web interface (at FNAL)**
- **Skyserver implement with SQL server and limited subset of CAS as a test (at JHU and FNAL)**
- **STScI hosts front-end with eventual transfer of all data distribution their.**
- ***Objectivity, STScI have been dropped; Skyserver has morphed into the CAS***



Plan 1

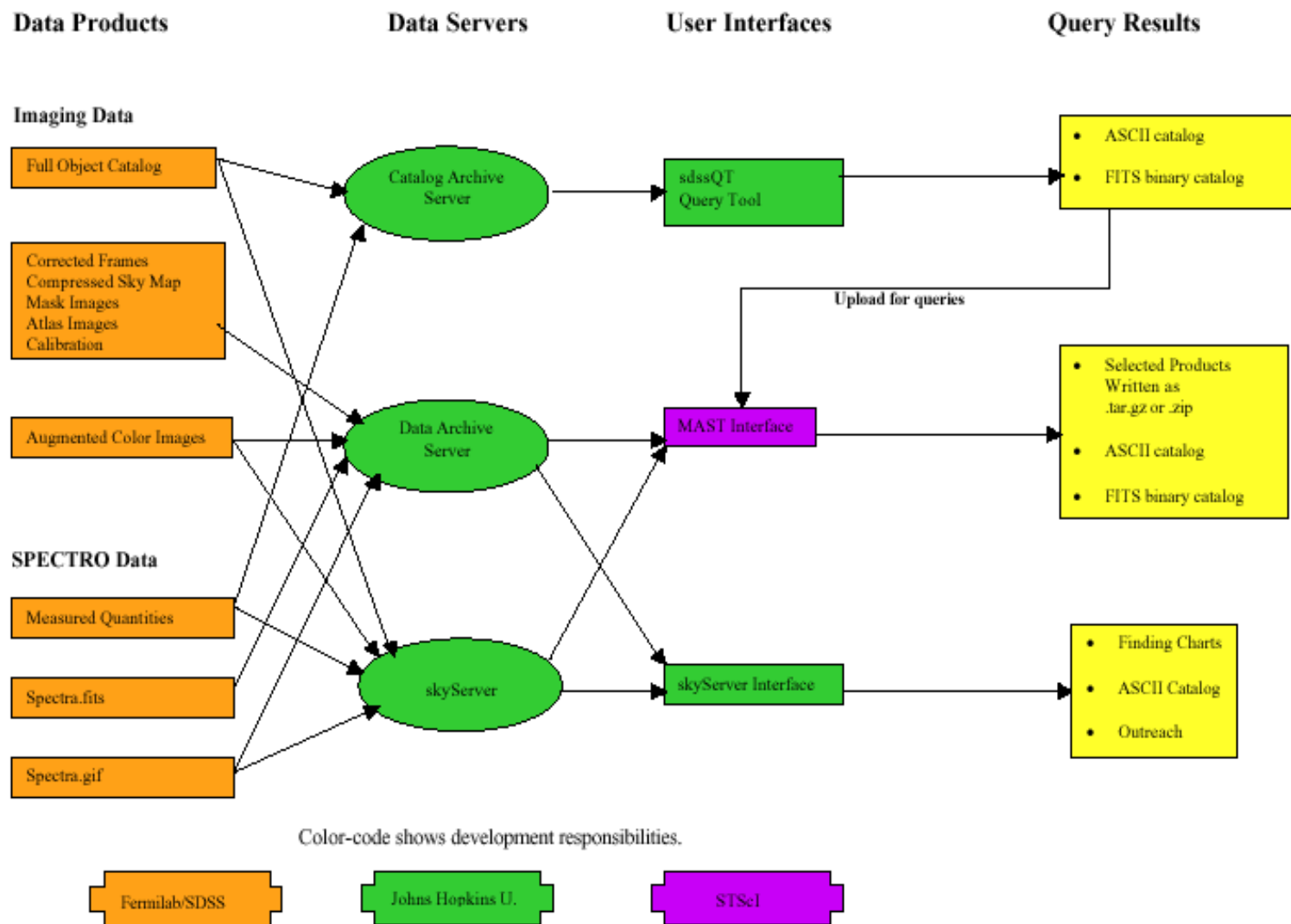


Figure 1. Data Flow Chart



- **Plan 2 (DR1)**
 - **DAS implemented with more sophisticated web interface (FNAL)**
 - **MySQL database backend with subset of CAS (FNAL)**
 - **CAS (a.k.a. Skyserver) delivered later than DAS (FNAL, now at JHU)**



• Plan 3

- **DAS as in plan 2 but with some web interfaces moved to CAS (FNAL)**
- **CAS as in plan 2 (FNAL, JHU)**